The Knowledge Engineering Review, Vol. 19:3, 187–212. © 2005, Cambridge University Press
doi:10.1017/S0269888905000251 Printed in the United Kingdom

An overview of methods and tools for ontology learning from texts

ASUNCIÓN GÓMEZ-PÉREZ and DAVID MANZANO-MACHO

Abstract

Ontology learning aims at reducing the time and efforts in the ontology development process. In recent years, several methods and tools have been proposed to speed up this process using different sources of information and different techniques. In this paper, we have reviewed 13 methods and 14 tools for semi-automatically building ontologies from texts and their relationships with the techniques each method follows. The methods have been grouped according to the main techniques followed and three groups have been identified: one based on linguistics, one on statistics, and one on machine learning. Regarding the tools, the criterion for grouping them, which has been the main aim of the tool, is to distinguish what elements of the ontology can be learned with each tool. According to this, we have identified three kinds of tools: tools for learning relations, tools for learning new concepts, and assisting tools for building up taxonomies.

1 Introduction

The Semantic Web has marked another stage in the ontology field. According to Berners-Lee (1999), the Semantic Web is an extension of the current Web where information is given well-defined meaning to better enable computers and people to work in cooperation. This cooperation can be achieved by using shared knowledge components, and thus ontologies have become key instruments. For these reasons, the ontological engineering community is exploring new methods and techniques to reduce the time and effort needed in the knowledge acquisition process, thus facilitating the construction of the new ontologies to be used by the emergent Semantic Web applications.

An ontology (Studer *et al.*, 1998) is 'a formal, explicit specification of a shared conceptualization'. A 'conceptualization' refers to an abstract model of some phenomenon in the world formed by identifying the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are clearly defined. 'Formal' refers to the fact that the ontology should be machine understandable. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group. Acquiring domain knowledge requires much time and many resources (i.e. human efforts, the analysis of several knowledge sources, etc.). Although ontology engineering tools have matured over the last decade, manual ontology acquisition remains a tedious, cumbersome task that can easily result in a knowledge acquisition bottleneck (Maedche & Staab, 2001).

Recently, the ontological engineering community has tried to discover new methods and tools for speeding up the ontology development process. The *ontology learning* process is defined as the application of a set of methods and techniques used for building an ontology from scratch by enriching, or adapting, an existing ontology in a semi-automatic fashion using distributed and heterogeneous knowledge and information sources, allowing a reduction in the time and effort needed in the ontology development process. The process includes a number of complementary disciplines applied to different types of unstructured, semi-structured, and fully structured data to support semi-automatic, cooperative ontology engineering (Maedche & Staab, 2004).

The ontology learning field brings a number of research activities together which focus on different types of knowledge and information sources but which share their target of a common domain conceptualization. Ontology learning is a complex, multi-disciplinary field that uses natural language processing (NLP), machine learning and knowledge representation. This paper attempts to cast some light on these issues by describing the most relevant methodological and technological approaches proposed for building a new ontology semi-automatically, paying special attention to learning from text approaches.

In this paper, the main approaches to ontology learning from natural language text are presented giving special attention to the main techniques that each approach follows for acquiring new knowledge. Also, the most relevant tools developed for helping the ontologist in the acquisition process are summarized. Out of the scope of this paper is the benchmarking of the ontology learning tools. Right now, there is no framework corpus that allows us to measure and experiment with the performance against a given set of standards for each technique and technology developed in the ontology learning field. In Section 2, we explore the most relevant approaches presented in the last few years; they are classified according to the type of source used in the learning process. In Section 3, we summarize the most relevant methods for ontology learning from natural language texts, identifying their goals and scope, the main techniques used in the process, and the steps followed in the learning process. In Section 4, we present the main tools developed to give technological support to these methods, if they exist. Finally, in Section 5, we present and discuss the conclusions extracted from this overview.

2 Approaches for ontology learning

Ontology learning has been brought about to help ontology engineers to construct ontologies and it is focused on the knowledge acquisition task. Though the fully automatic acquisition of knowledge remains distant, in this paper we take the view that the overall process is considered as semi-automatic, meaning by this that human intervention is necessary in some parts of the learning process. In the last decade several approaches have appeared for the partial automatization of the knowledge acquisition process. To carry out this automatization, NLP and ML techniques can be used. In this sense, Alexander Maedche and Steffen Staab (2004) distinguish the following:

- **Ontology learning methods from unstructured sources** consist in developing ontologies by applying natural language analysis techniques to texts. This group of approaches has a close relation with the NLP community, and it generally uses the information obtained from linguistic annotation processes over a selected corpus.[1] The linguistic annotation shows different levels of information (Aguado-de Cea *et al.*, 2002)—lemmatic, morphological, syntactic, semantic and discursive—this information being of great use in the learning process. Through this information, new concepts and the relations among them are learned by analysing the interactions and constraints given through the text. Another important source of support comes from the information extraction and text mining field which allow the detection and extraction of relevant information from texts.
- **Ontology learning from semi-structured sources** involves eliciting an ontology from sources that have some predefined structure, such as XML schemas. The new standards for document publishing on the Web have allowed the proliferation of semi-structured data while formal descriptions of semi-structured data are freely and widely available.
- **Ontology learning from structured data** aims at building an ontology extracting relevant concepts and relations from knowledge contained in structured data, such as databases. Ontologies have

[1] A corpus of texts is 'a set of texts that should be representative of the domain (complete), prepared to be processed by a computer, and accepted by the domain experts'. Definition extracted from Enery and Wilson (2001).

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

> been used for mediating between different databases. Ontology learning researchers aim to overcome the problem of building these ontologies manually by using learning methods.

In this article, we present an overview of the most relevant approaches that use text as the main input for developing an ontology semi-automatically and the respective technologies reported in the literature.

3 Methods for ontology learning from texts

This section discusses different methods and tools for ontology learning from texts. All the methods presented here use selected texts as the main input for learning an ontology or for enriching an existing ontology in a specific domain. Some of these approaches, however, complement the information extracted from the text with other domain knowledge sources, such as dictionaries, lexicons, term glossaries, etc. The methods based on text have a strong relationship with NLP, since they use annotation tools to extract linguistic information from the selected corpus, though they use different techniques to manage the texts and extract the ontology.

Below we present, in alphabetical order, the most relevant methods and approaches reported in the literature in the last few years that have been proposed for learning ontologies using texts. We have grouped these methods according to the main technique used for discovering new knowledge relevant for the domain. We have therefore distinguished between methods that apply linguistic techniques, those that apply statistical approaches, and those that apply machine learning algorithms. This classification, however, is arbitrary in the sense that all methods usually combine several techniques to achieve their goals and thus they could belong to more than one category. Since the main information source used to learn is text, all the methods need to apply some form of linguistic processing technique to manage text and to extract pieces of relevant information to the target domain. Nevertheless, we have classified the methods according to the technique that, in our opinion, has been applied to the different methods, and which allows us to distinguish and compare them.

Table 1 summarizes all the methods presented in this section according to the following criteria: the main goal of the method, the main techniques applied for learning, the possibility of reusing an existing ontology, the information sources used for learning, the availability of a tool that gives technological support to the method, and the process proposed for evaluating the final results.

3.1 Approaches based on linguistic techniques

This group encompasses all those methods that mainly base their operation on linguistic techniques, including, for example, linguistic patterns, pattern-based extraction, semantic relativeness measures, etc. These techniques are highly correlated with the kind of linguistic processing required by the method. Thus, linguistic patterns and several relativeness measures are generally constructed based on morphological or syntactic features detected in the text. Moreover, this information is also applied together with positional information referring to where an element involved in a pattern appears inside the text and what elements are close to it. The linguistic-based methods are commonly applied together with some statistical approaches to determine the relevance of each element found out in the textual sources to the target domain. The most relevant approaches in this group are summarized as follows.

Alfonseca & Manandhar (2002a,b) propose a new approach for extending existing ontologies such as WordNet (Miller, 1995); this approach is based on acquiring contextual properties of the words that co-occur with each set of concepts, following a top-down classification algorithm. It can then be used either to cluster concepts inside an ontology or to refine the ontology by adding new concepts. This approach is based on the hypothesis of Distributional Semantics (Alfonseca & Manandhar, 2002a): 'The meaning of a word is highly correlated to the contexts in which it

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Table 1 Summary of methods for ontology learning from text

Name of method	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation
Agirre *et al.*'s method (2000)	To enrich concepts in existing ontologies Control over sense proliferation	Statistical approach Topic signatures	WordNet	Domain Text	Information not available in the literature	User
Alfonseca and Manandhar's method (2002a,b)	To enrich an existing ontology with new concepts	Semantic relativeness Topic signatures	WordNet	Domain text WordNet	Welkin	Expert
Aussenac-Gilles *et al.*'s method (2002a,b)	To learn concepts and relations between them	Linguistic and semantic patterns Statistical approach	Domain ontology	Selected domain text	TERMINAE	User and expert
Faatz and Steinmetz's method (2002)	To enrich an existing ontology with new concepts	Semantic relativeness Statistical approach	Domain ontology	Domain text	Any ontology workbench	Expert
Hahn *et al.*'s method (1998, 2001)	To learn new concepts	Statistical approach Concept hypothesis based on quality labels	No	Domain text	Information not available in the literature	Empirical measures and by an expert
Hearst's method (1992)	To create a thesaurus, and also to enrich WordNet with new lexical–syntactic relations	Linguistic patterns	WordNet	Text WordNet	Information not available in the literature	Expert, comparing the results with WordNet
Hwang's method (1999)	To elicit a taxonomy	Linguistic patterns Statistical approach	No	Domain text	Information not available in the literature	Expert

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Table 1 *Continued*

Name of method	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation
Khan and Luo's method (2002)	To learn concepts	Machine learning algorithm for clustering Statistical approach	WordNet	Domain text	Information not available in the literature	Expert
Kietz *et al*.'s method (2000)	To learn concepts and relations between them to enrich an existing ontology	Text-mining algorithm to find non-taxonomical relation Statistical approach	Domain ontologies	Domain and non-domain specific Text	Text-To-Onto	User
Missikoff *et al*.'s method (2002)	To build taxonomies and to merge with an existing ontology	Machine learning techniques to identify semantic relations Statistical approach	WordNet	Domain text	OntoLearn	Expert
Moldovan and Girju's method (2000)	To enrich an existing ontology	Linguistic patterns	WordNet	Domain text Lexical resources	Information not available in the literature	User
Roux *et al*.'s method (2000)	To enrich a taxonomy with new concepts	Linguistic verb patterns	Domain ontology	Domain text	Information not available in the literature	Expert
Xu *et al*.'s method (2002)	To learn concepts and relations between them	Linguistic patterns using text-mining Statistical approach	WordNet	Annotated text corpus WordNet	Information not available in the literature	Expert

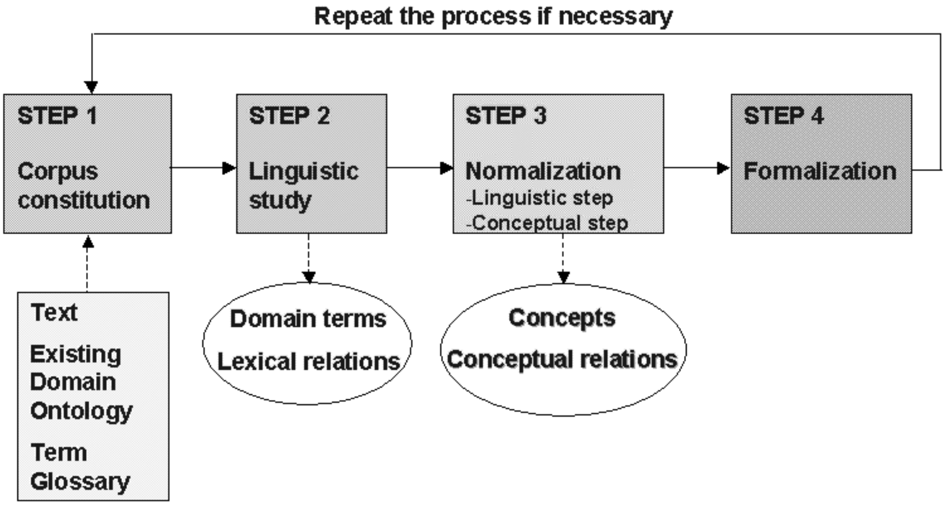

Figure 1 Steps proposed in Aussenac-Gilles *et al.*'s method

appears'. The contexts can be encoded as vectors of context words, as in the case of topic signatures (Lin & Hovy, 2000). Using topic signatures, each concept is represented by the set of words that co-occur with it and the frequencies with which they appear. Various similarity metrics, such as TF.IDF (Salton, 1991), can be used to measure the distance between the different concepts. The quality of the topic signatures can be improved by including only those context words that have some syntactic relationship with the concepts in the ontology. As the method is based on contextual information, it requires that several occurrences of the concepts to be classified are available, so that there is enough contextual information to generate the topic signatures. This approach receives technological support from the Welkin tool (Alfonseca & Manandhar, 2002c).

Aussenac-Gilles *et al.*'s method (2000a,b) allows the creation of a domain model by analysing a corpus with NLP tools and linguistics techniques, thus helping the ontologist to build up the ontology. This method combines knowledge acquisition tools based on linguistics with modelling techniques that allow links between models and source texts. The method uses texts and may use existing ontologies or terminological resources to build the ontology. The overall process is described in Figure 1. The activities proposed in the method are described below.

1. *Corpus constitution*. Texts are selected from the available domain-specific documentation according to the ontology requirements. The corpus has to cover the entire target domain and the selection of texts should be made by an expert of the domain. To perform this activity it is very useful to have a glossary of domain terms. Thus, the expert selects texts containing the terms of the glossary.
2. *Linguistic study*. This activity consists of selecting adequate linguistic tools and techniques and applying them to the texts. The main difficulty is in selecting the tools to be used, since they strongly depend on the language to be processed. As a result of this activity, domain terms, lexical relations, and groups of synonyms will be obtained. The extraction of terms is based on a frequency analysis of term occurrences and the extraction of relations is performed by means of some linguistic patterns.
3. *Normalization*. This activity includes a linguistic phase and a conceptual modelling phase. During the *linguistic phase*, the designer has to select the terms and the lexical relations (hyperonym, hyponym, etc.) to be modelled. Also in this *phase*, the ontologist adds a natural language definition for these terms taking into account the senses they have in the source texts. If terms with several meanings exist, the most relevant to the domain are kept. During the *conceptual phase*, concepts and semantic relations are defined in a normalized form using the labels of the concepts and relations. The result has to be checked according to differentiation rules, which require that for any given concept, the following information should be made explicit in the model: the attribute or relation the concept has in common with its father concept, the specific attribute or relation that makes it different from its father concept, and the property

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

that makes it different from its brothers. The result of this activity is a conceptual model expressed with a semantic network.
4. *Formalization*. This activity includes ontology validation and implementation.

The tools that give support to the different steps of this method are LEXTER (Bourigault *et al*. 1996), an NLP tool for terminology extraction; GEDITERM (Aussenac-Gilles, 1999), to define, model and consult a terminology connected to a semantic network; Caméléon (Aussenac-Gilles & Seguela, 2000), to extract relations; and TERMINAE (Biébow & Szulman, 1999) which can be used as a modelling tool.

Hahn *et al*. present a method for the maintenance (1998) and growth (2001) of domain-specific taxonomies based on natural language text understanding. A given taxonomy is incrementally updated as new concepts are acquired from real-world texts. The acquisition process is focused on the linguistic and conceptual 'quality' of various forms of evidence, such as generation and refinement of concept hypotheses. These concept hypotheses are ranked according to credibility, thus the most credible ones are selected for assimilation into the domain ontology. In this approach, learning is achieved by the refinement of multiple hypotheses about the concept membership of an instance. New concepts are acquired taking two sources of evidence into account: background knowledge from the domain texts, and linguistic patterns in which unknown lexical items occur. The model presented for text-based knowledge elicitation can be summarized to consist of the following general steps.

1. *Language processing*. This step aims at determining structural dependency information from the grammatical constructions in which an unknown lexical item occurs. The conceptual interpretation of these structures involving unknown lexical items in the terminological knowledge base is used to derive concept hypotheses.
2. *Calculation of the quality labels*. There are two kinds of quality labels to be calculated. The first is the *linguistic quality labels* that reflect structural properties of phrasal patterns or discourse contexts in which unknown lexical items occur and, depending on the type of the syntactic construction, different hypothesis generation rules may fire. The second type is the *conceptual quality labels* that result from comparing the representation structures of a concept hypothesis with those of alternative concept hypotheses or representation structures already existing in the underlying domain knowledge base. These quality labels are further enriched by conceptual annotations that reflect structural patterns of consistency, analogy, etc. This kind of initial evidence is represented by the corresponding sets of linguistic and conceptual quality labels.
3. *Quality estimation*. The overall credibility of single concept hypotheses is estimated by considering the available set of quality labels for each hypothesis calculated in the previous step. These quality labels are ranked in a list according to a preference order for the entire set of hypotheses. Whenever new evidence for or against a concept hypothesis is brought about, all concept hypotheses are re-evaluated.
4. *Evaluation*. An empirical evaluation of the text knowledge acquisition process is performed using different measures that evaluate the learning accuracy and the learning rate. The learning is achieved by the refinement of multiple hypotheses about the concept membership.

Hearst's method (1992) aims at acquiring automatically hyponym lexical relations from a corpus in order to build up a general domain thesaurus, using WordNet (Miller, 1995) to verify and augment its performance. The process uses a set of predefined lexico-syntactic patterns that are easily recognizable. These patterns occur frequently and across text genre boundaries indicating the lexical relation of interest. The method aims to discover instances of these patterns and can be also used to acquire other lexical relations. All these patterns will be used to build up the thesaurus, though they can be useful for other purposes, such as lexicon augmentation or semantic relatedness information. Hearst proposes the following procedure to discover new patterns automatically.

1. *Decide on a lexical relation that is of interest*. In this case, this is a subset of the hyponymy relation.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

2. *Gather a list of terms for which this relation is known to hold.* This list can be found automatically, bootstrapping from patterns found by hand, or bootstrapping from an existing lexicon or knowledge base.
3. *Find a place in the corpus* where these expressions occur syntactically near one another and record the environment.
4. *Find the commonalities* among these environments and hypotheses, of which the most commonyield patterns that indicate the relation of interest.
5. Once a new pattern has been positively identified, it is used to *gather more instances of the target relation* and go to step 2. To validate the resulting patterns, Hearst proposes comparing the results with the information found in WordNet (Miller, 1995). In this comparison, three kinds of outcomes are possible (Hearst, 1998):
   (a) *to verify*: if the two terms presented in the hyponymy relation are in WordNet, and if the relation between them is in the hierarchy, the thesaurus is verified;
   (b) *to criticize*: if the two terms presented in the new relation are in WordNet, but the relation is not in the hierarchy, the thesaurus is criticized, and a new set of hyponym connections is suggested to be added to WordNet;
   (c) *to augment*: if one or both terms presented in the new relation are not present, these noun phrases and their relationships are suggested as new entries to WordNet.

Moldovan & Girju's method (2000) allows discovering domain-specific concepts and relationships in an attempt to extend an existing ontology, like WordNet (Miller, 1995), with new knowledge acquired from parsed text. The source for discovering new knowledge is a non-domain-specific corpus, augmented by using other lexical resources such as domain-specific and general dictionaries. The user provides a number of domain-specific concepts that are used as seed concepts and she/he performs the validation of the correctness of the new concepts and relations learnt. This method can also be applied to learn an ontology from machine-readable dictionaries. It is composed of the following five steps.

1. *Selecting seed concepts.* Some seed concepts, which any user may consider important for the target domain ontology, are selected. This set of seed-concepts is extended with each concept's corresponding synonyms to form a synset. The knowledge to be acquired has to be related to one or more of these seed concepts, and consists of new concepts not defined in the existing ontology as well as new relations. The new relations link the new concepts with other concepts, some of which may already be present in the existing ontology.
2. *Discovering new concepts.* To discover new concepts from a general corpus, the method comprises the following phases (Moldovan & Girju, 2001). First, *documents that contain seed concepts are retrieved* and stored before being processed. The method only considers the nouns as candidate concepts. Second, for each document, *sentences that contain the seed concepts are extracted.* Only the noun phrases are considered. Third, *each of the previous sentences are part of speech tagged and parsed.* Finally, after parsing all sentences, *new concepts are extracted* (Modica *et al.*, 2001).
3. *Discovering lexical—syntactic patterns.* The aim of this step is to discover semantic relations between concepts (between two new concepts or between a new concept and one present in the existing ontology). The method then needs to use a new corpus provided by the user and different from the corpus used in the previous step. New noun sentences are extracted from this corpus. The objective here is to search for lexico-syntactical patterns comprising the concepts of interest, extracted in the previous step, inside the new group of sentences.
4. *Discovering new relations between concepts.* To carry out this process, three elements are used: the new concepts discovered in Step 2, the group of noun sentences extracted in that step, and the lexical-syntactic patterns resulting from Step 3. For each new concept the process tries to find all of the syntactic relations established in Step 3 in which the concept is involved. The relation is created between the two concepts linked by the syntactic relation. The validation of the process is performed by the user.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

5. *Classifying and integrating*. In this step a new taxonomy is created for the newly acquired concepts. This new taxonomy will be integrated with the existing taxonomy using the relations discovered in the previous step between a new concept and other concepts in the existing ontology (Harabagiu & Moldovan, 2000).

Roux *et al.*'s approach (2000) aims at enriching an existing ontology with new concepts extracted from a parsed domain corpus using NLP techniques. The approach is based on conceptual graphs (Sowa, 1984) and the idea behind it is to use the syntactic dependencies, focused on verb patterns extracted at the linguistic level, to build a semantic representation. According to these verb patterns, concepts are added to the ontology. However, when this approach is applied, two restrictions appear. First, the concept can only be an expression or a proper noun. Second, the data in the text should be easily identified by their immediate context. Roux *et al.* propose a general step to achieve these goals. When a new word that has not yet been referenced appears in the text as a concept in the existing ontology, it is necessary to add this new word as a new concept. As the ontology comprises concepts connected to each other along semantic paths, it is necessary to classify this new concept to find its correct place in the ontology. This approach is focused on managing new concepts with certain verb patterns that will define their position in the ontology. The verb patterns used in the approach are graphs where one of the nodes is a verb. The verb and the other elements included in the graph have to have certain semantic attributes. These graphs will serve to connect the new term that matches in a specific semantic context under its corresponding place in the ontology.

3.2 Approaches based on statistical techniques

This group includes all those methods that are mainly based on calculating several statistical measures which help the ontologist detect new concepts or relations between them. Among these techniques, frequency analysis of word repetition (or patterns of words) and TFIDF are usually applied mostly to score the relevance of the discovered elements for the target domain ontology. These techniques are usually applied with other techniques based on linguistics.

Agirre *et al.*'s method (2000) is aimed at enriching the concepts in existing large ontologies and controlling the proliferation of meanings inside them; it uses texts retrieved from the World Wide Web. The overall goal of this approach is to overcome two shortcomings of large ontologies such as WordNet (Miller, 1995): the lack of topical links among concepts, and the proliferation of different senses for each concept. This approach is based on *topic signatures* (Lin & Hovy, 2000), used in text summarization, and it follows the next four steps to improve an existing ontology with new concepts and relations between them.

1. *Retrieving relevant documents related to each concept in the ontology from the Web*. Queries are constructed for each concept sense with information contained in the ontology, such as synonyms of the concept, hyperonyms, attributes, etc. The documents that may be related to more than one sense are discarded, and documents related to the same concept sense are grouped together to form collections, one for each sense.
2. *Building topic signatures*. The documents in each collection that are related to a specific concept sense should be processed in order to extract words and their frequencies using a statistical approach, and the words most closely related to the concept are collected. Then, the data from one collection are compared with the data in the others. The words with a distinctive frequency in a collection are grouped in a list ordered by word frequency; this list constitutes the topic signature for each concept sense.
3. *Clustering word senses*. Given a word, the concepts that lexicalize its word sense are hierarchically clustered. To carry out this task different topic signatures are compared to discover shared words and to determine overlaps between the signatures. Various semantic relativeness metrics and clustering methods can be used for this purpose.
4. *Evaluating*. This is performed by the user comparing the results with a word sense disambiguation algorithm.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Faatz & Steinmetz (2002) present an approach the aims of which are to enrich an existing ontology by extracting meaning from the World Wide Web and to compare statistical information of word usage in a corpus with the structure of the ontology itself. Each concept in the ontology should have one or more phrases or words in natural language associated with it. Using this information, the approach proposes a method to calculate the similarities between words in order to enrich the concept definition and to create clusters of words related to a new concept. The new concepts will be proposed to a domain expert who will decide whether to add them to the ontology. This approach proposes the following general steps in order to create new concepts.

1. *Corpus constitution*. The source used for learning is a special corpus of text derived from the World Wide Web search results.
2. *Detection of a set of candidate concepts from the corpus*. The core idea of this step is to compute enrichment rules that do not contradict the semantic distance information already given by the ontology to be enriched. The corpus is statistically analysed, and a list of co-occurrences is generated for each word in the corpus. New words, related to the descriptors of each concept, are extracted based on a semantic distance function. These words, or their possible synsets, will be candidates to be new concepts.
3. *Selection of a subset of candidate concepts*. The list created in the previous step is proposed to a domain expert who will decide if they are relevant or not for the domain, and if they should be integrated into the ontology.

Xu *et al*.'s approach (2002) is designed to acquire domain-relevant terms and their relations using unsupervised hybrid text-mining techniques. This approach is based on the use of two different text-mining techniques to learn lexico-syntactic patterns, such as near synonymy relations, which indicate domain-relevant syntactic relations between the extracted terms. The first technique uses an existing ontology as initial knowledge for learning lexico-syntactic patterns, while the second is based on different co-location acquisition methods to deal with the free word-order language. The input for the process consists of a collection of pre-classified and linguistically annotated documents. In summary, the approach makes use of language parsing, an existing general ontology and statistical measures. The overall process can be explained as follows. First, it is necessary to *mine relevant terms* for the domain, for which several measures based on categorized documents are applied. Next, the process aims to *learn relations with lexico-syntactic patterns* between the terms extracted from the corpus, using the relations contained in WordNet (Miller, 1995) to assign synonymy, hyponymy and meronymy relations. To perform this activity, text fragments containing these semantic relations are extracted and similar relations are grouped to build clusters of patterns. At the end of this process, two types of patterns can be identified: domain-specific patterns, that define reliable domain-specific relations; and domain independent patterns. With these grouped relations, and with the extracted terms, *clusters of terms* can be created. Finally, with a learning term co-location activity, terms are put into the correct place *in the taxonomy* using the patterns mentioned before, and with statistical measures calculated for each pattern.

3.3 Approaches based on machine learning algorithms

This group includes all those methods that use several learning algorithms to assist the ontologist in detecting new concepts or relations between them, and to help find their correct place in the taxonomy. Machine learning research has bred a number of automated techniques for knowledge capturing and revision (Fayyad *et al*., 1996). Researchers on knowledge acquisition have looked for integrative approaches that exploit synergies between traditional knowledge acquisition approaches combined with machine learning techniques in order to discover, capture, represent, store, retrieve and reuse knowledge. Machine learning offers a set of techniques, tools and systems that can help to develop techniques and principles for automating acquisition of knowledge (Mitchell, 1997). In this section, we present some examples of this combination applied in the ontology learning field. These techniques are usually applied together with other techniques, mainly those from linguistics.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Hwang's approach (1999) aims at representing and retrieving information from large textual databases. This approach is based on the use of dynamic ontologies that capture the semantics of information present in the documents. The ontology is organized in simple taxonomies, and concepts from the taxonomy are then identified within the documents to enable the retrieval process. To carry out the learning process, NLP and machine learning techniques are used. The procedure for generating the ontology includes the following steps.

1. *Human experts provide* the system with a small number of *seed-words* that represent high-level concepts. Relevant documents will be collected from the Web automatically (with part of speech tagged or otherwise unmarked text).
2. The system *processes the incoming documents*, extracts only those phrases that contain seed words, generates corresponding concept terms, places them in the 'right' place in the ontology, and alerts the human experts to the changes. This feature is named 'discover-and-alert'. At the same time, it collects candidates for seed words for the next round of processing. The iteration then continues for a predefined number of times. The method indexes documents for future retrieval according to the concepts identified within them. It also indexes the 'context lines' in which the concept has been discovered to show how the concept was used in the text as well as the frequency of co-occurrence inside each document.
3. Several kinds of *relations* are extracted. Examples of relations are: 'is-a', 'part-of', 'manufactured-by', 'owned-by', etc., which are extracted according to linguistic features. The 'assoc-with' relation is used to define all relations that are not an 'is-a' relation. The method can only discover some of the attributes associated with certain concepts based on linguistic characters.

In each iteration, a *human expert is consulted* to ascertain the correctness of the concepts. If necessary, the expert has the right to make the correction and reconstruct the ontology.

Khan & Luo's method (2002) was developed to build a domain ontology from text documents using clustering techniques and WordNet (Miller, 1995). The method constructs the ontology in a bottom-up fashion. First, it constructs a hierarchy using some clustering techniques. Documents similar in content are associated with the same concept in the ontology. Next, a concept for each cluster of documents relative to the same topic in the hierarchy is assigned using a bottom-up concept assignment mechanism. To achieve this goal, a topic tracking algorithm (Joachims, 1997) and WordNet are used. This method consists of the following steps.

1. *Selection of the corpus to be used.* The user provides a selection of documents concerning the same domain.
2. *The construction of a hierarchy*. Using the set of documents created in the previous step, Khan and Luo aim to create a set of clusters where each cluster may contain more than one document, and put them into the correct place in a hierarchy. Each node in this hierarchy is a cluster of documents.
3. *Assignment of a concept*. After building a hierarchy of clusters, a concept is assigned to each cluster in the hierarchy using a bottom-up fashion. First, concepts associated with documents are assigned to leaf nodes in the hierarchy. For each cluster of documents a keyword is assigned which is called the topic; this topic represents its content and uses predefined topic categories. Then, the topic is associated with an appropriate concept in WordNet. Finally, the interior node concepts are assigned according to the concepts in the descendant nodes and their hypernyms in WordNet. If there is a relation between concepts in the hierarchy it is ignored, it is known only that there is a relation between them.

Kietz *et al.*'s method (2000) (see Maedche & Staab, 2001) is a generic method to discover a domain ontology from given heterogeneous resources using natural language analysis techniques. It is a semi-automatic process in the sense that the user takes part in the process. In their approach, Kietz *et al.* have adopted balanced cooperative modelling (Morik, 1993), where the work of building the ontology is distributed between several learning algorithms and the user. The method is based on the assumption that most concepts and conceptual structures of the domain to be

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

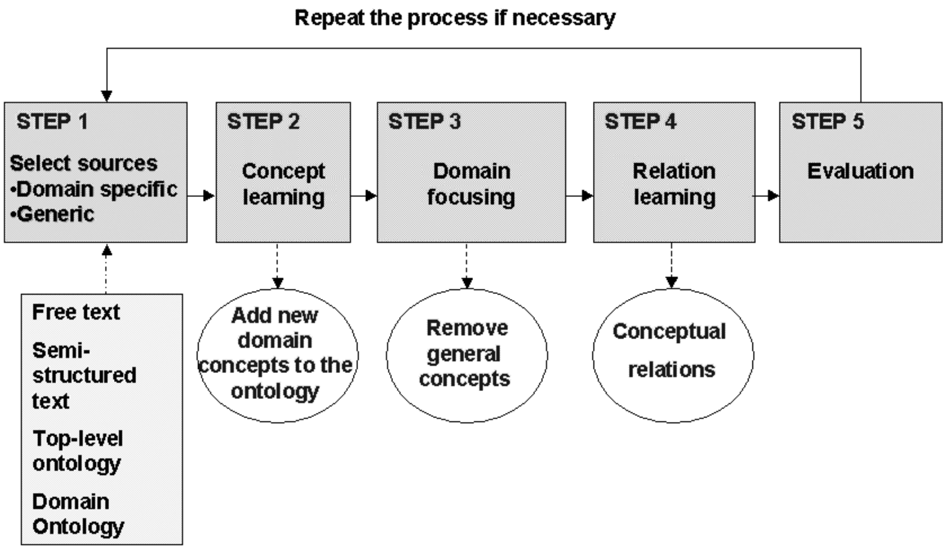

Figure 2 Steps followed in Kietz *et al.*'s method

included in an ontology and the terminology of a given domain are described in documents. The process is cyclic in the sense that the resulting ontology can be refined by applying the method iteratively. The acquisition process, summarized in Figure 2, consists of the following steps.

1. *Source selecting*. The process starts with the selection of a generic (top-level) ontology, which is used as a base in the learning process. This ontology should contain generic and domain concepts. The user must specify which documents should be used in the steps that follow to refine and extend the previous ontology. By their own nature, sources are heterogeneous in their formats and contents. They can be free text documents, semi-structured text, domain text, and non-domain-specific text. On the other hand, documents can be general or domain specific.
2. *Concept learning*. The goal of this step is to acquire new generic and specific concepts in order to decide if the discovered concepts are specific enough to be included in the ontology. The method involves analysing the frequency of the terms. Those terms that are more frequent in a domain-specific corpus than in generic corpora (and which are not contained in the given ontology) should be proposed to the user to decide whether they should be incorporated into the ontology.
3. *Domain focusing*. This step is aimed to prune the enriched core ontology by removing general concepts.
4. *Relation learning*. Frequency analysis can be used to learn ad hoc relations of the domain. This is founded in the underlying idea that frequent couplings of concepts in sentences can be considered as relevant relations between concepts in the ontology. This approach is used to find frequent correlations between concepts and it is based on the association rules algorithm proposed by Srikant and Agrawal (1995).
5. *Evaluating*. The resulting ontology has to be evaluated and the user has to decide whether it is necessary to repeat the process again.

The method is technologically supported by the Text-To-Onto (Maedche & Volz, 2001) tool.

Missikoff *et al.* (2002) (see Navigli *et al.*, 2003) have developed a method for ontology construction and enrichment using NLP and machine learning techniques. The method proposes using WordNet (Miller, 1995) as a source of prior knowledge to build a core domain ontology, after pruning all of the non-domain-specific concepts. The approaches followed by the method are statistical, to determine the relevance of one term for the domain, and semantic interpretation, based on machine learning techniques, to identify the right sense of terms and the semantic relations between them. Missikoff *et al.* propose three main steps to achieve their goals (Velardi *et al.*, 2002): terminology extraction, semantic interpretation, and creation of a specialized view of WordNet.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

1. *Terminology extraction*. Terms and combinations of terms, such as '*last week*', are extracted from a parsed corpus using NLP techniques. Terms are considered as the surface appearance of relevant domain concepts. High frequency in a corpus is a property observable for terminological as well as non-terminological expressions. The method uses a measure of the specificity of a terminology candidate with respect to the target domain via comparative analysis across different corpora. For this purpose, two different elements are defined to determine a threshold for the relevance of one terminology expression for the domain. The first element is the *domain relevance score*, which is a measure of the amount of information captured in the target corpus related to the entire collection of corpora used in the learning process. The second element is the *domain consensus*, which captures those terms that appear frequently across documents of a given domain.
2. *Semantic interpretation*. The main goals of this step are to determine the right concept sense for each component of a complex term, as in a semantic disambiguation process, and then to identify the semantic relations holding among the concepts to build a complex concept. At the end of this step, a domain concept forest will be obtained, showing the taxonomic and other relationships among complex domain concepts represented by expressions. To carry out this step, it is necessary to use semantic and linguistic resources (the method has been tested with WordNet) to assist in the semantic interpretation of terms. This step consists of two main processes, the first of which is a *semantic disambiguation process*. The sense of each word is defined as a synset of synonyms (i.e. the WordNet synset into the word can be correctly placed). The second process involves extracting semantic relations that hold between the components of complex terms extracted in the previous step.
3. *Creating the domain ontology*. This step is intended to integrate the taxonomy obtained in the previous step with a core domain ontology. In case an existing domain ontology is not available, Missikof *et al*. propose to create a new one from WordNet, pruning concepts not related to the domain, and extending the newly created domain ontology with the new domain concept trees under the appropriate nodes.

This method is partially supported by the OntoLearn (Velardi *et al*., 2001) tool.

3.4 Comparison of ontology learning methods from texts

In this section, we have presented several approaches for learning ontologies from texts. All the approaches apply different linguistic techniques and statistical or machine learning techniques, except for the methods of Hearst, Moldoban & Girju, and Roux *et al*., which are mainly linguistics based. These approaches can be also compared taking into account whether the method provides the possibility of reusing other existing ontologies to improve the learning process. In this sense, WordNet is the most common ontology reused of the methods reviewed here, and it is used as an initial ontology which can be enriched with new concepts or relations. Because this ontology stores general linguistic information, it is easy to use in combination with natural language analysis techniques, and can be used in different ways inside each learning method. For example, Agirre *et al*. and Alfonseca & Manandhar use WordNet and enrich it with new concepts; other methods, such as those of Missikoff *et al*. and Xu *et al*. use WordNet as a lexicon to detect synonyms and other linguistic relations from the selected texts. Other methods that reuse a different ontology take an existing domain ontology with the aim of enriching it with new concepts and relations. This is the case for the methods of Faatz & Steinmetz and Roux *et al*. Kietz *et al*.'s method is the only one that allows a general domain to be used as an input, with the aim of extracting a domain ontology skeleton by means of removing unspecific domain concepts with pruning techniques.

One clear result arises from this overview, and it is the need for an expert to evaluate the resulting ontology after the learning process.

According to the previous analysis, we can conclude that:

(a) a detailed methodology or method that guides the ontology learning process does not exist and, regardless of the approach considered, there are only methods that provide general guidelines;

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Table 2 Summary of ontology learning tools

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability
ASIUM	To learn taxonomic relations	Conceptual clustering techniques	Factorization Clustering	Text syntactically analysed	Whole process	Can be used to perform the knowledge acquisition in any other ontology development
Caméléon	To tune generic lexico-syntactic patterns or build new ones, and to find taxonomic and non-taxonomic lexical relations	To reuse and tune generic patterns, Hearst's proposal, and pattern identification to help learn lexical relations	Own method	Texts processed by taggers Its own base of generic patterns	Validates, adapts, or defines new domain-specific patterns and relations. Domain expert just validates the model	Imports lists of terms from any term extractor
LTG	To discover internal relations of texts in natural language	Statistical approach Linguistic patterns	Own method	Plain text	Whole process	Can be used to perform the knowledge acquisition in other ontology development tools
Mo'K Workbench	To assist in concept formation	Conceptual clustering	Own method	Tagged text	Whole process	Can be used to perform the knowledge acquisition in other ontology development tools
OntoLearn	To help detect new concepts and relations to be added to an existing ontology	Statistical approach Machine learning techniques for pattern discovery	Missikoff *et al.*'s method	Plain text	Evaluation	Information not available in the literature

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Table 2 *Continued*

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability
Prométhée	To extract and refine lexical–syntactic patterns	Learning from examples	Own method	Pattern bases	Whole process	Information not available in the literature
SOAT	To discover new relations	Linguistic patterns	Own method	Plain text	Information not available in the literature	Information not available in the literature
SubWordNet Engineering Process	To acquire new lexical concepts	Linguistic patterns Statistical approaches	Own method	Plain text	Whole process	Information not available in the literature
SVETLAN'	To build clusters of nouns	Conceptual clustering	Own method	Plain text	Not required	Information not available in the literature
TERMINAE	To learn new concepts	Conceptual clustering	Own method	Plain text	Validation	Information not available in the literature
Text-To-Onto	To find taxonomic and non-taxonomic relations, and to assist in concept formation	Association rules Statistical approach	Kietz *et al*.'s method	Plain text Ontologies	Whole process	KAON tool suite
TextStorm	To detect new relations	Linguistic patterns	Own method	Plain text	Whole process	Information not available in the literature
Welkin	To enrich existing general ontologies with new terms	Semantic similarity measures	Alfonseca & Manandhar's method	Plain text WordNet	Validation	None
WOLFIE	To learn a semantic lexicon	Statistical approach	Own method	Pre-processed plain text	Validation	Information not available in the literature

(b) a complete correspondence between the approaches for ontology learning and the tools developed does not exist; the tools give only partial technological support to perform some of the steps proposed in the different approaches, except in Kietz *et al.*'s method and its corresponding tool.

The linguistic-based methods are highly dependent on the type of NLP performed to analyse the textual sources and the quality of the sources selected for the learning process. In this sense, they are more difficult to apply to other domains or types of texts. The statistical-based methods present the advantage of being easy to implement. Nevertheless, the methods which only make use of statistical approaches have some disadvantages: since they consider the information inside the text as symbols, they do not make use of the meaning that words have inside the text. A word has a different meaning if it goes together with another word (i.e. an adjective specifies the meaning of a noun, an adverb modifies the sense of the action expressed by the verb, etc.). For this reason, most of the methods summarized in this paper combine the statistics with NLP. The machine-learning-based approaches make use of machine learning algorithms to help the ontologist find out specific elements of the ontology (i.e. Kietz *et al.* make use of data mining algorithms to detect non-taxonomic relations) or to reduce the complexity of the data to be presented to the ontologist, who will decide upon the kind of element and whether it is relevant enough to the target domain.

4 Tools for ontology learning from texts

In this section, we present the most relevant tools for ontology learning from texts. These tools aim to provide support for the knowledge acquisition process and for the subsequent ontology learning. Not all the methods presented in the previous sections have an associated tool which gives technological support to achieve their goals, and some tools do not follow any of methods reviewed. We have grouped the tools presented in this section into three types: tools for learning relations, tools for learning new concepts, and tools for assisting in building a taxonomy. The first group includes those tools that help to detect new relations (taxonomic or non-taxonomic) from the selected input. The second group covers all those tools that assist the ontologist in finding and setting up new concepts. Finally, the last group deals with the tools that help the ontologist build a taxonomy or enrich an existing one. All the tools in each group are presented in alphabetical order.

Table 2 summarizes some conclusions about the tools presented in this section. We have compared them with the same evaluation criteria, and these include their main goal and scope, the main technique followed, the method used for learning, the sources, the user intervention in the process, and their interoperability with other ontology development tools.

4.1 Tools for learning relations

This section presents a set of tools that helps the ontologist find out new relationships (taxonomic or non-taxonomic) among concepts to enrich an existing ontology. These tools are based mainly on statistical approaches combined with linguistic information to determine the existence and the type of the new relation.

ASIUM (Faure & Nédellec, 1999; Faure & Poibeau, 2000) is an acronym of 'Acquisition of Semantic knowledge Using Machine learning methods'. The main aim of ASIUM is to help the expert in the acquisition of semantic knowledge and taxonomic relations among terms, extracted from technical texts using syntactic analysis. ASIUM takes as input French texts in natural language and it associates a frequency of occurrence to each word in the text. The learning method is based on conceptual and hierarchical clustering. Basic clusters are formed by words that occur with the same verb after the same preposition (Faure & Nédellec, 1998). The tool uses a metric to compute the semantic similarity between clusters; this metric is then used by the ontologist to decide

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

if a new concept is created. Clusters are successively aggregated by the conceptual clustering method to form the concepts of the ontology. Thus the ontologist defines a minimum threshold for gathering clusters into concepts, and then the learning is validated by the ontologist. The tool follows two steps to achieve its results. The first is *factorization* (conceptualization): the head words are associated with their frequency of appearance in the text to calculate the distances among concepts. Those that appear in similar contexts are added, by means of an algorithm of conceptual clustering, to form the concepts of the ontology. For this purpose, a technique to estimate the semantic similarity among concepts has been used (Bisson, 1992a,b; Liu, 1996). The second step is *clustering* (ontology building). Since a hierarchy would be too restricted to represent the complexity of the ontology in many domains, the authors have adopted the technique of pyramidal clustering, and thus the ontology is constructed level by level.

Caméléon (Aussenac-Gilles & Seguela, 2000) has been developed to assist in learning conceptual relations to enrich conceptual models. Caméléon relies on linguistic principles for relation identification (i.e. lexico-syntactic patterns are good indicators of semantic relations). Some patterns may be regular enough to indicate the same kind of relation. Other patterns are domain specific and may reveal domain-specific relations. Learning conceptual relations with Caméléon is a two-fold process. The first part is dedicated to the identification of the relevant patterns and relations for the current corpus. Generic patterns from a generic base are available in the tool; they must be evaluated for the current corpus, and they may be modified or rejected. New specific patterns may be identified either manually or using Hearst's principle (1992) with pairs of domain-specific related terms. The second part is dedicated to the use of these patterns to identify lexical relations and to manually enrich a conceptual model from them. Patterns are used to list all possible lexical relations in the texts; their evaluation provides suggestions of conceptual relations. For each concept in the model, lexical relations are presented and must be validated to enrich the model. Finally, this tool gives technological support to some steps of Aussenac-Gilles *et al.*'s method.

The LTG (Language Technology Group) Text Processing Workbench (Mikheev & Finch, 1997) is a set of computational tools for uncovering internal structure in natural language texts written in English. The main idea behind the workbench is the independence of the text representation and text analysis. In LTG, ontology learning is performed in two sequential steps: representation and analysis. In the representation step, the text is converted from a sequence of characters to features of interest by means of annotation tools. In the analysis step, those features are used by tools for statistics gathering and inference to find significant correlations in the texts. The analysis tools are independent of any particular assumption about the nature of the feature set and work on the abstract level of feature elements, which are represented as SGML items. The workbench is being used both for lexicographic purposes and for statistical language modelling. It supports an incremental process of corpus analysis starting from a rough automatic extraction and organization of lexical—semantic regularities and ending with a computer-supported analysis of extracted data and a semi-automatic refinement of obtained hypotheses. To do this, the workbench uses methods from computational linguistics, information retrieval and knowledge engineering.

Prométhée (Morin, 1998, 1999) is a machine-learning-based tool for extracting and refining lexical—syntactic patterns related to concept relations from technical corpora. It uses pattern bases, which are enriched with the pattern bases extracted during learning. To refine patterns, the authors propose the Eagle (Martienne & Quafafou, 1998) learning system. This system is based on the inductive paradigm *learning from examples* (Mitchell, 1997), which consists of extracting intentional descriptions of target concepts from their extensional descriptions and previous knowledge of the given domain. Eagle extracts *intentional* descriptions of concepts from their *extensional* descriptions. The learned definitions are later used in recognition and classification tasks. The interface between the two systems works as follows: first, Prométhée extracts lexical—syntactic patterns; then, some instances of these patterns are produced from the corpus and classified among the examples of the patterns; and, finally, from these labelled patterns Eagle produces descriptions that are interpreted as restrictions refining the patterns.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

The SOAT tool (Wu *et al*., 2002) allows semi-automatic domain ontology acquisition from a domain corpus. The main objective of the tool is to extract relationships from parsed sentences based on applying phrase rules to identify keywords with strong semantic links such as hyperonyms or synonyms. The acquisition process is based on the use of InfoMap (Hsu *et al*., 2001), a knowledge representation framework that integrates linguistic, commonsense, and domain knowledge. InfoMap has been developed to perform natural language understanding and to capture the topic words, usually pairs consisting of a noun and a verb, or two nouns, in a sentence. InfoMap has two major types of relations between concepts: taxonomic relations (category and synonym) and non-taxonomic (attribute and event). The acquisition process carried out by SOAT includes collecting domain keywords and finding the relationships between them. To perform this activity, a set of rules has been defined to extract keywords from a sentence; this sentence is related to concepts in InfoMap that have a strong semantic relation between them. The tool receives as input a domain corpus with tags from parts of speech. A keyword, usually the name of the domain, is selected as a root in the corpus. Then, with this keyword, the process aims to find a new keyword related to the previous one by applying extraction rules and adding the new keyword to the ontology, following the rules and the structure fixed in InfoMap. This new keyword is now taken as a root to repeat the process for a pre-determined number of times or until it is impossible to find a new related keyword. User intervention is necessary to verify the results of the acquisition and to refine and update the extraction rules. The restrictions of SOAT impose that the quality of the corpus should be very high, in the sense that the sentences must be accurate and that there should be enough sentences including most of the important relationships to be extracted.

4.2 Tools for learning new concepts

The tools reported in this section aim to help the ontology to form new concepts or to enrich, in the case of linguistic ontologies such as WordNet, the content of the synsets with new elements. Such tools mainly base their performance on implementing different clustering techniques following machine learning algorithms or linguistic-based approaches.

Mo'K Workbench (Bisson *et al*., 2000) is a configurable workbench that supports the semiautomatic construction of ontologies from a corpus using different conceptual clustering methods. Mo'K assists ontologists in the exploratory process of defining the most suitable learning method. In this sense, Mo'K supports the elaboration, comparison, characterization and evaluation of different conceptual clustering methods. It also permits fine-grained definitions of similarity measures and class construction operators, easing the task of method instantiation and configuration. The learning process supported by this workbench takes a corpus as input. No additional knowledge is used to label the input, to guide the learning, or to validate the learned results. Through NLP techniques, the tool extracts a list of triplets from the corpus. A triplet is composed of a verb, a word and a syntactic role of this word in a sentence. Using the triplets, Mo'K calculates the number of occurrences of each. Triplets with a low number of occurrences or too many occurrences are removed from the list. Finally, Mo'K calculates the semantic distance between the triplets of the list to form conceptual clusters.

The SubWordNet Engineering Process tool (Gupta *et al*., 2002) acquires and maintains sublanguage WordNets. The architecture builds upon WordNet's (Miller, 1995) semantic structure and includes integrated capabilities for concept element discovery, concept identification, and concept maintenance. The architecture to perform each of these capabilities has been modularized into two layers: the concept discovery layer, and the data layer. The *concept discovery capability* includes the Concept Discovery Workbench, a Concept Discovery Engine, and Discovered Concepts Database modules. This layer provides a GUI that allows users to select the documents, to manipulate discovered concept elements, and to provide summary distributional information of words and phrases. The layer also includes several NLP components to discover relations using collocation statistics, lexical patterns, etc. The *concept identification capability* includes the

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Concept Identification Workbench, the Concept Identification Engine, and the Identified Concepts Database modules. These capabilities have been designed to support concept identification and phrase clustering and to establish concordances between concept nodes and WordNet synsets.

SVETLAN' (Chaelandar & Grau, 2000) is a domain-independent tool that creates clusters from words appearing in texts. Its learning method is based on a distributional approach: nouns playing the same syntactic role in sentences with the same verb are aggregated in the same class. The learning process has the following steps: syntactic analysis, aggregation and filtering. In the syntactic analysis step, the tool retrieves sentences from the original texts in order to find the verb inside the sentence. This is based on the assumption that verbs can be used for categorizing nouns. The output of this step is a list of triplets that contain a verb, a noun and the syntactic relation between them. The aggregation step constructs groups of nouns with similar meanings. The filtering step is based on the weight of the nouns inside their classes. It removes nouns from these groups if they are not very relevant for the class. The threshold is established by the ontology developer. The process does not require validation and is completely independent of the ontology developer.

Welkin (Alfonseca & Manandhar, 2002b,c) is a tool for generating e-learning materials from unrestricted texts automatically; there is one particular module of the architecture that tries to create an ontological representation of the terms of interest appearing in the text, to represent internally the different sections in the e-learning web sites. The aim of this module is to analyse the texts in order to identify relevant terminology and to classify those terms inside lexical ontologies such as WordNet (Miller, 1995). For classifying, contextual information is used. Each of the concepts in the original ontology is extended with information about which specific words can appear in their contexts, and which of those have syntactic relationships with each concept. A distance metric, based upon the representations of the contexts, is then used to classify the new terms inside the ontology. The final part of the procedure is performed by a module that looks for word patterns expressing relationships between the concepts. Welkin gives support to carry out Alfonseca and Manandhar's method.

WOLFIE (WOrd Learning From Interpreted Examples) (Thompson & Mooney, 1997) learns a semantic lexicon from a corpus. The lexicon learned consists of words paired with representations of their meaning, and it allows both synonymy and polysemy. WOLFIE is part of an integrated system that learns to parse novel sentences into their meaning representations. The system combines the following features: first, arbitrary amounts of both polysemy and synonymy can be handled; then, WOLFIE interacts with the system CHILL (Yamaguchi, 1999), which learns to parse database queries directly into logical form; and finally, the algorithm used for learning is fast and accurate, and deals with the best selection of phrase meanings based on several heuristics. The idea behind its algorithm is that each choice of a lexical item may constrain the possible meanings of phrases not yet learned. To achieve its goal, the system makes a few assumptions about the problem: the meaning of a sentence is composed of possible meanings of words and phrases in that sentence; the sentence representation contains no noise; and the meaning of each occurrence of a word in a sentence appears only once in the sentence's representation.

4.3 Tools that assist in building a taxonomy

In this section, we present the tools that assist the ontologist in the whole process, or at least in some of the tasks of building up an ontology (or enriching an existing one). These tools commonly combine several types of techniques and also interoperate with an NLP tool, which provides linguistic information from the selected information sources.

OntoLearn (Velardi *et al*., 2002) is aims to extract relevant domain terms from a corpus of text, to relate them to appropriate concepts in a general-purpose ontology, and to detect relations among the concepts. To perform these tasks, natural language analysis and machine learning techniques are used. The tool has been tested within the European Harmonise project. OntoLearn extracts terminology from a corpus of domain text such as specialized Web sites. The tool then filters the

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

terms using natural language processing and statistical techniques that perform comparative analysis across different domains, or it contrasts corpora. This analysis identifies terminology that is used in the target domain but not seen in other domains. Next, it uses the WordNet (Miller, 1995) lexical knowledge bases to perform semantic interpretation of the terms. The tool then relates concepts according to taxonomic (kind-of) and other semantic relations, generating a domain concept forest. To extract such relations, WordNet and a rule-based inductive-learning method have been used. Finally, OntoLearn integrates the domain concept forest with WordNet to create a pruned and specialized view of the domain ontology, and the validation of the process is performed by an expert. Finally, Ontolearn is used to perform Missikoff *et al.*'s method, giving technological support to perform some of the steps proposed in that method.

TERMINAE (Biébow & Szulman, 1999; Szulman *et al.*, 2002) integrates linguistic and knowledge engineering tools. The linguistic tool allows the definition of terminological forms from the analysis of term occurrences in a corpus. The ontologists analyse the uses of the term in the corpus to define the meanings of the terms. The knowledge engineering tool includes an editor and a browser for the ontology. TERMINAE helps represent terminological forms as a concept (called terminological concept). TERMINAE builds concepts from the study of the corresponding term in a corpus. First, the tool establishes the list of terms, and this requires the constitution of a relevant corpus on the domain. Using a term extractor tool, a set of candidate terms are proposed to the ontologist, who selects a set of terms. Then, the ontologist conceptualizes the terms and analyses the uses of the term in the corpus to define all the meanings of the term. The ontologist then gives a definition in NL for each meaning, and translates the definition into an implementation language. TERMINAE gives some technological support for Aussenac-Gilles *et al.*'s method.

Text-To-Onto (Maedche & Staab, 2000; Maedche & Volz, 2001) integrates an environment for building domain ontologies from an initial core ontology. It also discovers conceptual structures from different German sources using knowledge acquisition and machine learning techniques (Agrawal *et al.*, 1993). Text-To-Onto has implemented some techniques for ontology learning from free and semi-structured text. The result of the learning process is a domain ontology that contains domain-specific and domain-independent concepts. Domain-independent concepts are withdrawn to better adjust the vocabulary of the domain ontology. The result of this process is a domain ontology that only contains domain concepts learnt from the input sources mentioned before. The whole process is supervised by the ontologist. This is a cyclic process, in the sense that it is possible to refine and complete the ontology if the process is repeated. Finally, Text-To-Onto is the tool used to perform Kietz *et al.*'s method.

The TextStorm and Clouds framework has been developed within the Dr Divago project (Pereira, 1998) for constructing a semantic network semi-automatically that contains only concepts and their relations, using a relevant text for the target domain. It is composed of two main modules: TextStorm (Oliveira *et al.*, 2001), and Clouds (Pereira *et al.*, 2000), which perform complementary activities. TextStorm deals with the task of extracting relations between concepts from a text file using NLP techniques, while Clouds is concentrated on completing these relations and extrapolating rules of the knowledge previously extracted using NLP techniques. *TextStorm* is an NLP tool that extracts binary predicates from a text using syntactic and discourse knowledge. The process starts by providing the system with texts that contain relevant features of the target domain. Then, the text is tagged using a WordNet (Miller, 1995) database to find all the parts of speech to which a word may belong and to classify words in the parsing process. Next, the text is parsed using an augmented grammar to obtain a lexical classification of the words. The predicates on which the tool is focused are all those that relate two concepts in sentences. These predicates are only verbal phrases that contain two nouns (subject and direct object) connected with a verb that specifies an existent relation between them. With this information, TextStorm creates a list that will be the input for the Clouds tool. To perform the process, the system needs to interact with the user, who has to resolve inconsistencies and to decide the relevance of a sentence for the domain. *Clouds* is responsible for the construction of a semantic network in an interactive way. Using the previous list,

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

with all binary predicates extracted from the text, Clouds builds a hierarchical tree of concepts, and learns some particulars of the domain using two different techniques. The first one is called a *best current hypothesis based algorithm*, and is used to learn the categories of the arguments of each of the relations. The other is an *inductive logic programming based algorithm*, used to learn the contexts that are recurrent in each relation. To perform the process, Clouds will ask the user about new concepts and new relations that it suspects exist.

4.4 Comparison of ontology learning methods from texts

If we compare these tools according to their goals and scope, all of them assist in extracting useful domain knowledge from the selected text to form new concepts or to detect new relations among them. However, only a small group really help to build a whole taxonomy. This is the case, for example, of Text-To-Onto, TERMINAE or OntoLearn, all of which belong to the group of tools that assist in building up a taxonomy, and they are the only ones that allow other existing ontologies to be reused to speed up the learning process. Text-To-Onto and OntoLearn take an existing ontology to extract a core domain ontology using a pruning technique, while TERMINAE can take a domain ontology to be enriched with new concepts and new relations.

The tools presented use different *learning techniques* to reach their goals. In this sense, there are tools which use a linguistic patterns approach to discover new relations. This is the case of ASIUM, Caméléon, LTG, the SubWordNet Engineering Process, SOAT and TextStorm. Another set of tools use basically conceptual clustering techniques aiming to construct new concepts. MO'K, SVENTLAN', and TERMINAE belong to this group. Finally, there is another group which uses mainly statistical approaches combined with machine learning algorithms. Tools such as OntoLearn, Text-To-Onto, and Wolfie belong to this group. However, all the tools need natural language processing to some degree to perform their activities.

Concerning the *methods* to which the tools give support, only OntoLearn, TERMINAE, Text-To-Onto, and Welkin have methods associated to those presented in Section 3. OntoLearn follows Missikof *et al.*'s method; Terminae gives some technological support to Aussenac-Gilles *et al.*'s method; Text-To-Onto implements Kietz *et al.*'s approach, and Welking supports Alfonseca & Manandhar's method. The rest of the tools follow their own learning method.

As for *user intervention* during the learning process, all the tools have in common that their activities are performed by interacting with the user. Some of them, such as Caméléon, OntoLearn, TERMINAE, Welkin, and Wolfie, need the user to validate the results of the learning process. The others need the interaction with the user during the whole process. In this sense, all the tools lack validation techniques for the resulting learning, this task always being performed by the user.

Finally, taking into account the interoperability with other ontology development tools, only the Text-To-Onto tool has been integrated in the OntoEdit (Sure *et al.*, 2002) ontology development platform.

5 Conclusions

Ontology learning is a suitable process to accelerate the knowledge acquisition activity of the ontology development process. Ontology learning can be useful for building an ontology from scratch, reusing an existing one, or speeding up the construction of ontologies to be used for different purposes. However, the aim of building an ontology automatically is far from being achieved. From this study, we can extract the following conclusions.

From a *methodological point of view*, this review shows the following.

- A detailed methodology that guides the ontology learning process regardless of the source used for learning or the approach considered does not exist yet. There are only methods that provide general guidelines, and they need interaction with a user to achieve their goals.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

- There is not a complete correspondence between the approaches for ontology learning and the tools developed. These tools give only partial technological support to perform some steps proposed in the different approaches, except for Kietz *et al.*'s method and the Text-To-Onto tool.
- The methods presented for ontology learning from text are based mainly on natural language analysis techniques complemented with statistical measures. Such techniques are used to elicit new concepts or relations from the selected sources.
- The most common ontology included in the methods is WordNet, which is used as an initial ontology to be enriched with new concepts or relations. The reason for this is that this ontology stores general linguistic information, therefore it is easy to use with natural language analysis techniques.
- All these methods require the participation of an ontologist to evaluate both the final ontology and the accuracy of the learning process. There are no methods that evaluate the accuracy of the learning process.
- The application of some machine learning methods can help the KE reduce the time and cost of developing knowledge-based software by extracting knowledge directly from existing databases and textual repositories (Webb, 2002). Other machine learning methods enable software systems to improve their performance over time with minimal user intervention (Hastie *et al.*, 2001).

From a technological point of view, it is possible to reach the following conclusions.

- A fully automatic tool that carries out the learning process does not exist yet. Some tools are focused on helping with the acquisition of lexical-semantic knowledge, others assist in eliciting concepts or relations from a pre-processed corpus with the help of the user.
- None of the tools can evaluate the accuracy of the learning process or compare the different results obtained with different learning techniques, thus the participation of an ontologist is required to evaluate the final ontology.
- Further ontology learning works should carry out benchmarking studies of ontology learning tools measuring their performance against a standard or a given set of standards and making the comparison of similar processes in different contexts.[2] For this purpose, It will be necessary to establish a general framework where different techniques and technology could be compared and evaluated against the same standardized corpus, source ontologies and terminological resources. This kind of work would allow us to determine what technology is better for some kind of problems and to reuse technology already developed to solve other more complex problems.

Acknowledgements

This work has been partially supported by the IST thematic network OntoWeb (IST-2000–29243), the IST project Esperonto (IST-2001–34373), the CICYT project 'ContentWeb: Plataforma Tecnológica para la Web Semántica' (TIC-2001–2745) and a FPI (Formación de Personal Investigador) grant from CAM (Comunidad Autónoma de Madrid). We are also grateful to Rosario Plaza for her help in checking and reviewing the grammar and spelling of the paper and improving clarity and style.

References

Adriaans, P and Zantinge, D, 1996, *Data Mining* Reading, MA: Addison-Wesley.

Agirre, E, Ansa, O, Hovy, E and Martinez, D, 2000, ''Enriching very large ontologies using the WWW'' in S Staab, A Maedche, C Nedellec and P Wiemer-Hastings (eds.) *Workshop on Ontology Learning of the European Conference on Artificial Intelligence (ECAI'00) (CEUR Workshop Proceedings, 31)*, Berlin, pp. 25–30, http://CEUR-WS.org/Vol-31/.

[2] Deliverable 2.1.1. 'Survey of scalability techniques for reasoning with ontologies'. http://www.cs.vu.nl/~holger/KnowledgeWeb/Deliverables/D2.1.1/D2.1.1-StateOfTheArt.pdf

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Agrawal, R, Imielinski, T and Swami, A, 1993, ''Mining association rules between sets of items in large databases'' in P Buneman and S Jajodia (eds.) *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC*, pp. 207–216.

Aguado-de Cea, G, Álvarez de Mon-Rego, I, Gómez-Pérez, A, Pareja-Lora, A and Plaza-Arteche, R, 2002, ''A Semantic Web Page linguistic annotation model'' in *Semantic Web Meets Language Resources*. Technical report WS-02-16, Menlo Park, CA: AAAI Press, pp. 20–29.

Alfonseca, E and Manandhar, S, 2002a, ''An unsupervised method for general named entity recognition and automated concept discovery'' in *Proceedings of the 1st International Conference on General WordNet, Mysore, India*.

Alfonseca, E and Manandhar, S, 2002b, ''Improving an ontology refinement method with hyponymy patterns'' in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain, pp. 235–239.

Alfonseca, E and Manandhar, S, 2002c, ''Extending a lexical ontology by a combination of distributional semantics signatures'' in A Gómez-Pérez and VR Benjamins (eds.) *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02) (Lecture Notes in Artificial Intelligence, 2473)*. Berlin: Springer, pp. 1–7.

Aussenac-Gilles, N, 1999, ''Gediterm, un logiciel de gestion de bases de connaissances terminologiques'' *Terminologies Nouvelles* **19** 111–123.

Aussenac-Gilles, N, Biébow, B and Szulman, S, 2000a, ''Revisiting ontology design: A methodology based on corpus analysis'' in R Dieng and O Corby (eds.) *12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00) (Lecture Notes in Artificial Intelligence, 1937)*. Berlin: Springer, pp. 172–188.

Aussenac-Gilles, N, Biébow, B and Szulman, S, 2000b, ''Corpus analysis for conceptual modelling'' in N Aussenac-Gilles, B Biébow and S Szulman (eds.) *European Knowledge Acquisition Workshop 2000 (EKAW'00), Workshop on Ontologies and Texts, Juan-Les-Pins, France (CEUR Workshop Proceedings, 51)*, Amsterdam, pp. 1.1–1.8, `http://CEUR-WS.org/Vol-51/`.

Aussenac-Gilles, N, Biebow, B and Szulman, S, 2003, ''D'une méthode à un guide pratique de modélisation de connaissances à partir de textes'' in F Rousselot (ed.) *Rencontres Terminologie et IA (TIA 2003), Enssais*, pp. 41–53.

Aussenac-Gilles, N and Seguela, P, 2000, ''Les relations sémantiques: du linguistique au formel. Cahiers de grammaire. Numéro spécial Sur la Linguistique de Corpus'' in A Condamines (ed.) *Presse de l'UTM*, vol 25, Toulouse, pp. 175–198.

Berners-Lee, T, 1999, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. New York: HarperCollins.

Biébow, B and Szulman, S, 1999, ''TERMINAE: A linguistic-based tool for the building of a domain ontology'' in D Fensel and R Studer (eds.) *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and Management (EKAW'99), Dagstuhl, Germany (Lecture Notes in Artificial Intelligence, 1621)*. Berlin: Springer, pp. 49–66.

Bisson, G, 1992a, ''Learning in FOL with a similarity measure'' in P Rosenbloom and P Szolovits (eds.) *Proceedings of the 10th American Association for Artificial Intelligence conference*. Menlo Park, CA: AAAI Press.

Bisson, G, 1992b, ''Conceptual clustering in a first-order logic representation'' in B Neumann (ed.) *Proceedings of the 10th European Conference on Artificial Intelligence (ECAI'92)*. New York: John Wiley, pp. 458–462.

Bisson, G, Nedellec, C and Cañamero, D, 2000, ''Designing clustering methods for ontology building. The Mo'K Workbench'' in S Staab, A Maedche, C Nedellec and P Wiemer-Hastings (eds.) *Workshop on Ontology Learning of the European Conference on Artificial Intelligence (ECAI'00) (CEUR Workshop Proceedings, 31)*, Berlin, pp. 13–18, `http://CEUR-WS.org/Vol-31/`.

Bourigault, D, Gonzalez, I and Gros, C, 1996, ''LEXTER, a natural language tool for terminology extraction'' in *Proceedings of the Seventh EURALEX International Congress, Goteborg*.

Chaelandar, G and Grau, B, 2000, ''SVETLAN'—a system to classify words in context'' in S Staab, A Maedche, C Nedellec and P Wiemer-Hastings (eds.) *Workshop on Ontology Learning of the European Conference on Artificial Intelligence (ECAI'00) (CEUR Workshop Proceedings, 31)*, Berlin, pp, 19–24, `http://CEUR-WS.org/Vol-31/`.

Enery, TMC and Wilson. A, 2001, *Corpus Linguistics: An Introduction* Edinburgh: Edinburgh University Press.

Faatz, A and Steinmetz, R, 2002, ''Ontology enrichment with texts from the WWW'' in B Berendt, A Hotho and G Stumme (eds.) *Proceedings of the 2nd Workshop on Semantic Web Mining at 13th European Conference on Machine Learning (ECML'02), 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, Helsinki, pp. 20–35.

Faure, D and Nédellec, C, 1998, ''A corpus-based conceptual clustering method for verb frames and ontology acquisition'' in P Velardi. (ed.) *Adapting Lexical and Corpus Resources to Sublanguages and Application*

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Workshop of the 1st International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain, pp. 1–8.

Faure, D and Nédellec, C, 1999, ''Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM'' in D Fensel and R Studer (eds.) *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and Management (EKAW'99), Dagstuhl, Germany (Lecture Notes in Artificial Intelligence, 1621)*. Berlin: Springer, pp. 329–334.

Faure, D and Poibeau, T, 2000, ''First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX'' in S Staab, A Maedche, C Nedellec and P Wiemer-Hastings (eds.) *Workshop on Ontology Learning of the European Conference on Artificial Intelligence (ECAI'00) (CEUR Workshop Proceedings, 31)*, Berlin, pp. 7–12, `http://CEUR-WS.org/Vol-31/`.

Fayyad, U, Piatetsky-Shapiro, G and Smyth, P, 1996, ''From data mining to knowledge discovery: An overview'' in U Fayyad, G Piatetsky-Shapiro, P Smyth and R Uthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, pp. 1–36.

Gupta, KM, Aha, DW, Marsh, E and Maney, T, 2002, ''An architecture for engineering sublanguage WordNets'' in *Proceedings of the 1st International Conference on General WordNet (Mysore, India)*, pp. 207–215.

Hahn, U and Markó, K, 2001, ''Joint knowledge capture for grammars and ontologies'' in Y Gil, M Musen and J Shavlik (eds.) *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), Victoria, British Columbia*. New York: ACM Press, pp. 68–75.

Hahn, U and Schnattinger, K, 1998, ''Towards text knowledge engineering'' in *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI '98) and 10th Conference on Innovative Applications of Artificial Intelligence (IAAI'98), Madison, WI*. Menlo Park: AAAI Press/Cambridge, MA: MIT Press, pp. 524–531.

Hahn, U and Schulz, S, 2000, ''Towards very large terminological knowledge bases: A case study from medicine'' in HJ Hamilton (ed.) *Advances in Artificial Intelligence. Proceedings of the 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI'2000), Montreal (Lecture Notes in Artificial Intelligence, 1822)*. Berlin: Springer, pp. 176–186.

Hamp, B and Feldweg, H, 1997, ''GermaNet—a lexical—semantic net for German'' in Vossen, Calzolari, Adriaens, Sanfilippo and Wilks (eds.) *Proceedings of the Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications Workshop at the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL'97), Madrid.*

Hastie, T, Tibshirani, R and Friedman, J, 2001, *The Elements of Statistical Learning* Berlin: Springer.

Harabagiu, SM and Moldovan, DI, 2000, ''Enriching the WordNet taxonomy with contextual knowledge acquired from text'' in S Shapiro and L Iwanska (eds.) *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. Menlo Park: AAAI/Cambridge, MA: MIT Press, pp. 301–334.

Hearst, MA, 1992, ''Automatic acquisition of Hyponyms from large text corpora'' in Zampolli, A. (ed.) *Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France*, vol. 2. Morristown, NJ: Association for Computational Linguistics, pp. 539–545.

Hearst, MA, 1998, ''Automated discovery of WordNet relations'' in C Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, pp. 132–152.

Hovy, EH and Lin, C-Y, 1999, ''Automated text summarization in SUMMARIST'' in M Maybury and I Mani (eds.) *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, pp. 18–24.

Hsu, WL, Wu, SH and Chen, YS, 2001, ''Event identification based on the Information Map—InfoMap'' in *Natural Language Processing and Knowledge Engineering Symposium of the of the IEEE Systems, Man, and Cybernetics Conference, Tuckson, AZ*.

Hwang, CH, 1999, ''Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information'' in E Franconi and M Kifer (eds.) *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), Linköping, Sweden (CEUR Workshop Proceedings, 21)*, Berlin, pp. 14–20, `http://CEUR-WS.org/Vol-21/`.

Joachims, T, 1997, ''A probabilistic analysis of the Rocchio Algorithm with TFIDF for text categorization'' in DH Fisher (ed.) *Proceedings of 14th International Conference on Machine Learning (ICML-97), Nashville, TN*. San Francisco: Morgan Kaufmann, pp. 143–151.

Khan, L and Luo, F, 2002, ''Ontology construction for information selection'' in *Proceedings of 14th IEEE International Conference on Tools with Artificial Intelligence, Washington*, IEEE Computer Science Press, pp. 122–127.

Kietz, JU, Maedche, A and Volz, R, 2000, ''A method for semi-automatic ontology acquisition from a corporate Intranet'' in N Aussenac-Gilles, B Biébow and S Szulman (eds.) *European Knowledge Acquisition Workshop 2000 (EKAW'00), Workshop on Ontologies and Texts, Juan-Les-Pins, France (CEUR Workshop Proceedings, 51)*, Amsterdam, pp. 4.1–4.14, `http://CEUR-WS.org/Vol-51/`.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

Lin, C-Y and Hovy, EH, 2000, ''The automated acquisition of topic signatures for text summarization'' in M Kay (ed.) *Proceedings of the 17th Conference on Computational Linguistics, Strasbourg, France*, vol. 1. Morristown, NJ: Association for Computational Linguistics, pp. 495–501.

Liu, WZ, 1996, ''An integrated approach for different attribute types in nearest neighbour classification'' *The Knowledge Engineering Review* **11**(3) 245–252.

Maedche, A and Staab, S, 2004, ''Ontology learning'' in S Staab and R Studer (eds.) *HandBook on Ontologies (International Handbooks on Information Systems Series)*. Berlin: Springer, pp. 173–190.

Maedche, A and Staab, S, 2000, ''Discovering conceptual relations from text'' in W Horn (ed.) *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000), Berlin*. Amsterdam: IOS Press, pp. 321–325.

Maedche, A and Staab, S, 2001, ''Ontology learning for the semantic Web'' *IEEE Intelligent Systems, Special Issue on the Semantic Web* **16**(2) 72–79.

Maedche, A and Volz, R, 2001, ''Ontology extraction and maintenance environment the Text-To-Onto'' in FJ Kurfess and M Hilario (eds.) *Proceedings of the Integrating Data Mining and Knowledge Management Workshop at the IEEE International Conference on Data Mining, California*.

Martienne, E and Quafafou, M, 1998, ''Vagueness and data reduction in concept learning'' in H Prade (ed) *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98), Brighton*. Chichester: John Wiley and Sons, pp. 351–355.

Mikheev, A and Finch, S, 1997, ''A workbench for finding structure in texts'' in R Grishman (ed.) *Proceedings of the Applied Natural Language Processing (ANLP-97), Washington*. San Francisco, CA: Morgan Kaufmann, pp. 372–379.

Miller, GA, 1995, ''WordNet: A lexical database for English'' *Communications of the ACM* **38**(11) 39–41.

Missikoff, M, Navigli, R and Velardi, P, 2002, ''The usable ontology: An environment for building and assessing a domain ontology'' in I Horrocks and J Hendler (eds.) *Proceedings of the Semantic Web—ISWC 2002: First International Semantic Web Conference, Sardinia (Lecture Notes in Computer Science, 2342)*. Berlin: Springer, pp. 39–53.

Mitchell, T, 1997, *Machine Learning* New York: McGraw-Hill.

Modica, G, Gal, A and Jamil, HM, 2001, ''The use of machine-generated ontologies in dynamic information seeking'' in *Proceedings of the Sixth International Conference on Cooperative Information Systems (CoopIS 2001), Trento, Italy (Lecture Notes in Computer Science, 2172)*. Berlin: Springer, pp. 433–448.

Moldovan, DI and Girju, RC, 2000, ''Domain-specific knowledge acquisition and classification using WordNet'' in J Etheredge and B Manaris (eds.) *Proceedings of the Thirteenth International Florida Artificial Intelligence Research Symposium Conference, Orlando*. Menlo Park, CA: AAAI Press, pp. 224–229.

Moldovan, DI and Girju, RC, 2001, ''An interactive tool for the rapid development of knowledge bases'' *International Journal on Artificial Intelligence Tools* **10**(1–2) 65–86.

Moldovan, DI, Girju, RC and Rus, V, 2000, ''Domain-specific knowledge acquisition from text'' in S Nirenburg (ed.) *Proceedings of the Applied Natural Language Processing (ANLP-2000) Conference, Seattle*, pp. 268–275.

Morik, K, 1993, ''Balanced cooperative modelling'' *Journal of Machine Learning* **11**(1) 217–235.

Morin, E, 1998, ''Prométhée un outil d'aide a l'acquisition de relations semantiques entre temes'' in J Bouaud (ed.) *Actes of the 5th National Conference on Traitement Automatique des Langues Naturelles (TALN'98), Paris*, pp. 172–181.

Morin, E, 1999, ''Automatic acquisition of semantic relations between terms from technical corpora'' in *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering (TKE-99)*. Vienna: TermNet-Verlag, pp. 268–278.

Navigli, R, Velardi, P and Gangemi, A, 2003, ''Ontology learning and its application to automated terminology translation'' *IEEE Intelligent Systems* **18**(1).

Oliveira, A, Pereira, FC and Cardoso, A, 2001, ''Automatic reading and learning from text'' in R Akerkar (ed.) *Future Trends in Artificial Intelligence, Proceedings of ISAI'2001, Kolhapur, India*. New Delhi: Allied Publishers.

Pereira, FC, 1998, ''Modelling divergent production: A multi domain approach'' in H Prade (ed.) *13th European Conference on Artificial Intelligence (ECAI'98), Brighton, UK*. Chichester: John Wiley and Sons, pp. 131–132.

Pereira, FC, Oliveira, A and Cardoso, A, 2000, ''Extracting concept maps with clouds'' in G Henning (ed.) *Proceedings of the Argentine Symposium of Artificial Intelligence (ASAI 2000), Buenos Aires, Argentina*.

Rigau, G, Rodríguez, H and Agirre, E, 1998, ''Building accurate semantic taxonomies from monolingual MRDs'' in *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), Montreal*, vol. II, ICCL, Association for Computational Linguistics, pp. 1289–1293.

Roux, C, Proux, D, Rechermann, F and Julliard, L, 2000, ''An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions'' in S Staab, A Maedche, C Nedellec

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press

and P Wiemer-Hastings (eds.) *Workshop on Ontology Learning of the European Conference on Artificial Intelligence (ECAI'00) (CEUR Workshop Proceedings, 31)*, Berlin, pp. 49–51, `http://CEUR-WS.org/Vol-31/`.

Salton, G, 1991, ''Developments in automatic text retrieval'' *Science* **253** 974–979.

Seguela, P, 1999, ''Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés'' in *Actes de TIA'99 Terminologie et Intelligence Artificielle, Nantes, France (Terminologies Nouvelles, 19)*, pp. 52–60.

Sowa, JF, 1984, ''Conceptual structures'' in *Information Processing in Mind and Machine*. Reading MA: Addison-Wesley.

Srikant, R and Agrawal, R, 1995, ''Mining generalized association rules'' in U Dayal, P Gray and S Nishio (eds.) *Proceedings of 21th International Conference on Very Large Data Bases*. San Francisco: Morgan Kaufmann, pp. 407–419.

Studer, R, Benjamins, VR and Fensel, D, 1998, ''Knowledge engineering: Principles and methods'' *Data & Knowledge Engineering* **25** 161–197.

Sure, Y, Erdmann, M, Angele, J, Staab, S, Studer, R and Wenke, D, 2002, ''OntoEdit: Collaborative ontology engineering for the semantic Web'' in I Horrocks and J Hendler (eds.) *Proceedings of the Semantic Web—ISWC 2002: First International Semantic Web Conference, Sardinia (Lecture Notes in Computer Science, 2342)*. Berlin: Springer, pp. 221–235.

Szulman, S, Biebow, B and Aussenac-Gilles, N, 2002, ''Structuration de terminologies à l'aide d'outils d'analyse de textes avec TERMINAE'' *Traitement Automatique de la Langue (TAL). Numéro spécial sur le Structuration de Terminologie* **43**(1) 103–128.

Thompson, CA and Mooney, RJ, 1997, ''Semantic lexicon acquisition for learning parsers''. Technical note, January 1997.

Velardi, P, Missikoff, M and Fabriani, P, 2001, ''Using text processing techniques to automatically enrich a domain ontology'' in *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2001), Maine*. New York: ACM Press, pp. 270–284.

Velardi, P, Navigli, R and Missikoff, M, 2002, ''Integrated approach for Web ontology learning and engineering'' *IEEE Computer* **35**(11) 60–63.

Webb, GI, 2002, ''Integrating machine learning with knowledge acquisition'' in CT Leondes (ed.) *Expert Systems*, vol. 3. San Diego, CA: Academic Press, pp. 937–959.

Wu, SH and Hsu, WL, 2002, ''SOAT: A semi-automatic domain ontology acquisition tool from Chinese corpus'' in *Proceedings of the 19th International Conference on Computational Linguistic (COLING-ACL'02), Taipei*, ICCL, Association for Computational Linguistics, pp. 289–293.

Xu, F, Kurz, D, Piskorski, J and Schmeier, S, 2002, ''A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping'' in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.*

Yamaguchi, T, 1999, ''Constructing domain ontologies based on concept drift analysis'' in VR Benjamins, B Chandrasekaran, A Gomez Perez, N Guarino and M Uschold (eds.) *Proceedings of Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends (IJCAI-99), Stockholm.*

Zelle, JM, 1995, ''Using inductive logic programming to automate the construction of natural language parsers''. PhD dissertation, University of Texas, Austin, TX. Also appears as Artificial Intelligence Laboratory Technical report AI 96–249.

https://doi.org/10.1017/S0269888905000251 Published online by Cambridge University Press